**几个最新免费开源的中文语音数据集**

工欲善其事必先利其器，做机器学习，我们需要有利器，才能完成工作，数据就是我们最重要的利器之一。做中文语音识别，我们需要有对应的中文语音数据集，以帮助我们完成和不断优化改进项目。我们可能很难拿到成千上万小时的语音数据集，但是这里有一些免费开源的语音数据集，大家一定不要错过。文末附数据集下载地址。

**THCHS30**

THCHS30是一个很经典的中文语音数据集了，包含了1万余条语音文件，大约40小时的中文语音数据，内容以文章诗句为主，全部为女声。它是由清华大学语音与语言技术中心（CSLT）出版的开放式中文语音数据库。原创录音于2002年由朱晓燕教授在清华大学计算机科学系智能与系统重点实验室监督下进行，原名为“TCMSD”，代表“清华连续”普通话语音数据库’。13年后的出版由王东博士发起，并得到了朱晓燕教授的支持。他们希望为语音识别领域的新入门的研究人员提供玩具级别的数据库，因此，数据库对学术用户完全免费。

license: Apache License v.2.0

**ST-CMDS**

ST-CMDS是由一个AI数据公司发布的中文语音数据集，包含10万余条语音文件，大约100余小时的语音数据。数据内容以平时的网上语音聊天和智能语音控制语句为主，855个不同说话者，同时有男声和女声，适合多种场景下使用。

License: Creative Common BY-NC-ND 4.0 (Attribution-NonCommercial-NoDerivatives 4.0 International)

**AISHELL开源版**

AISHELL是由北京希尔公司发布的一个中文语音数据集，其中包含约178小时的开源版数据。该数据集包含400个来自中国不同地区、具有不同的口音的人的声音。录音是在安静的室内环境中使用高保真麦克风进行录音，并采样降至16kHz。通过专业的语音注释和严格的质量检查，手动转录准确率达到95％以上。该数据免费供学术使用。他们希望为语音识别领域的新研究人员提供适量的数据。

License: Apache License v.2.0

**Primewords Chinese Corpus Set 1**

Primewords包含了大约100小时的中文语音数据，这个免费的中文普通话语料库由上海普力信息技术有限公司发布。语料库由296名母语为英语的智能手机录制。转录准确度大于98％，置信水平为95％，学术用途免费。抄本和话语之间的映射以JSON格式给出。

**aidatatang**

Aidatatang\_200zh是由北京数据科技有限公司（数据堂）提供的开放式中文普通话电话语音库。

语料库长达200小时(当前时长最长的中文开源语音数据集)，由Android系统手机（16kHz，16位）和iOS系统手机（16kHz，16位）记录。邀请来自中国不同重点区域的600名演讲者参加录音，录音是在安静的室内环境或环境中进行，其中包含不影响语音识别的背景噪音。参与者的性别和年龄均匀分布。语料库的语言材料是设计为音素均衡的口语句子。每个句子的手动转录准确率大于98％。

**附下载链接：**

* THCHS30

[国内镜像](https://link.ailemon.me/?target=http://cn-mirror.openslr.org/resources/18/data_thchs30.tgz) | [国外镜像](https://link.ailemon.me/?target=http://www.openslr.org/resources/18/data_thchs30.tgz)

* ST-CMDS

[国内镜像](https://link.ailemon.me/?target=http://cn-mirror.openslr.org/resources/38/ST-CMDS-20170001_1-OS.tar.gz) | [国外镜像](https://link.ailemon.me/?target=http://www.openslr.org/resources/38/ST-CMDS-20170001_1-OS.tar.gz)

* AISHELL开源版

[国内镜像](https://link.ailemon.me/?target=http://cn-mirror.openslr.org/resources/33/data_aishell.tgz) | [国外镜像](https://link.ailemon.me/?target=http://www.openslr.org/resources/33/data_aishell.tgz)

* Primewords Chinese Corpus Set 1

[国内镜像](https://link.ailemon.me/?target=http://cn-mirror.openslr.org/resources/47/primewords_md_2018_set1.tar.gz) | [国外镜像](https://link.ailemon.me/?target=http://www.openslr.org/resources/47/primewords_md_2018_set1.tar.gz)

* aidatatang

[国内镜像](https://link.ailemon.me/?target=http://cn-mirror.openslr.org/resources/62/200.zip) | [国外镜像](https://link.ailemon.me/?target=http://www.openslr.org/resources/62/200.zip)